電気学会論文誌 C（電子・情報・システム部門誌）
IEEJ Transactions on Electronics, Information and Systems
Vol.135 No.9 pp.1136–1141 DOI: 10.1541/ieejeiss.135.1136

Paper

Autonomous Learning and Recognition of Human Action based on An Incremental Approach of Clustering

Wee-Hong Ong[*a)] Non-member, Leon Palafox[**] Non-member
Takafumi Koseki[*] Member

(Manuscript received July 15, 2014, revised March 18, 2015)

At current stage, the majority of the human activity recognition (HAR) technologies are based on supervised learning, where there are labeled data to train an expert system. In this paper, we proposed a framework based on the unsupervised learning to autonomously discover, learn and recognize atomic activities, i.e., the actions. The input to the HAR framework is a sample pool of unlabeled observations of an unknown number of actions. An incremental action discovery algorithm based on K-means is used to discover new actions. For each new action discovered, a learning algorithm is used to model it through an automated training and cross-validation cycle. The algorithm uses Mixture of Gaussians Hidden Markov Model (HMM) to model the actions, and the algorithm autonomously determines the appropriate number of Gaussian components and states. The framework deals with the dynamic and noisy nature of the data. We evaluated the proposed framework on a third party dataset of daily activities and the results show its performance is in-par with that achieved using a supervised learning algorithm to recognize the activities from the same dataset.

Keywords: action discovery;action modeling;action recognition;unsupervised learning;

1. Introduction

Due to the need of labeled data and supervised approaches, the generalization of Human Activity Recognition (HAR) technologies in regular human living environments is limited. In regular human living environments, observation data are not labeled. For human activity recognition technologies to be deployed in such environments, we require unsupervised approach.

While there are research works on unsupervised human activity recognition, they are either focused on solving computer vision problems using an unsupervised approach, require visual data from high cost sensors, address a specific aspect of human activity recognition or they require an alternative form of pre-labeled data from wearable sensors or other sources.

In this paper, we proposed a framework to enable the complete process of human activity recognition without using labeled data. The framework, at current stage, has been developed to deal with atomic activities. In our context, actions are atomic activities that do not decompose into meaningful sub-activities. They are the building blocks for higher level activities.

Two contributions arise from this paper:

First, we proposed a framework to autonomously discover, learn and recognize human actions using an unsupervised approach. The framework only requires unlabeled observations that are available in abundance in our regular living environment. It uses data from inexpensive consumer sensor allowing it to be implemented at low cost. The input to the framework is unlabeled observations of an unknown number of actions including random and unintentional movements. The framework is designed to reject these random movements.

Second, we demonstrated that it can perform well using an incremental approach of clustering, and a simple probability-based model, Hidden Markov Model (HMM). We proposed strategies to autonomously determine the required parameters for the HMM. Our empirical results suggest suitable choice of parameter values to keep computation time low.

The learned action models do not have a linguistic label. Each model is an action and enumerated. For the purpose of understanding human activities and intention, for example, it is sufcient for the intelligent system to know that Action 1 usually takes place in the morning, at the bed and that it usually leads to Action 2. It does not matter whether Action 1 is being labeled as wake up or “okimasu (wake up in Japanese language)” and Action 2 is being labeled as brush teeth or “hawomigaku (brushing teeth in Japanese language)”, or in other languages.

Linguistic labels are for the purpose of communication. This can be achieved through interaction with human; just like the way children label what they have learned in the environment by asking other people. The proposed framework in this paper is providing the ability similar to that the children discover and learn their environment, however without the final phase of linguistic labeling.

a) Correspondence to: Wee-Hong Ong. E-mail: weehong.ong@ubd.edu.bn
* Graduate School of Engineering, The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
** Department of Radiology, University of California
Los Angeles, CA 90095-7437

The remaining of this paper is organized as follows: In Section 2, we will describe related works. We describe our proposed framework in Section 3 with details of each phase. Section 4 describes the experiment conducted to demonstrate the effectiveness of the framework and the dataset used in the experiment. We report the results and discuss their implications in Section 5. Finally, we make conclusions in Section 6.

2. Related Works

HAR technologies can be broadly divided into two categories: sensor-based and vision-based. Given that our goal is to deploy HAR technologies in normal homes without embedded sensors, we have focused on vision-based HAR technologies. Turaga *et al.* [1], Poppe [2] and Aggarwal *et al.* [3] have made extensive survey and review of vision-based HAR technologies. From their reviews, we note all the works described in their papers have used supervised approach to learn the activity models.

In the normal home setting, labeled data are scarce and requiring user annotation of actions in daily activities is impractical. These conditions call for the requirement to perform HAR from unlabeled data. In our earlier paper [4], we have developed an incremental approach of clustering for human activity discovery from unlabeled data. Activity discovery refers to the process in which an intelligent agent finds new activities from its observations. In a typical activity recognition process, an intelligent agent will recognize known activities from its observations, i.e. it sees that the person is performing activity X. In activity discovery, the intelligent agent finds new activities in the observations that are not recognized as any known activity, i.e. it sees the person is performing a new activity.

Activity discovery is only an initial stage in HAR. The output from the activity discovery need to be used for learning and recognition of activities. In this paper, we build on top of the activity discovery algorithm to provide a complete HAR framework that includes other stages such as learning and recognition.

In this paper, we have also refined the scope of our work to use the term action instead of activity. We define actions as atomic activities that do not decompose into meaningful subactivities. For example, the cooking activity may include the actions of chopping and stirring. They are the building blocks for higher level activities. We will therefore be considering action discovery, learning and recognition.

When a new activity or action is discovered, i.e., a cluster is found, members of the cluster become examples to learn a template or model for that action. Probability-based algorithms are most commonly used to model human activities and actions [5]. In fact, Hidden Markov Model (HMM) [6] is one of the most popular models [1][5] to build human activity or action models. HMM has well established mathematical ground and is versatile in its application. It has been used to model human actions since the early 90s [7]. It remains effective even at present time [8] when the objective is for activity or action recognition but not novelty in learning algorithm. We have chosen Hidden Markov Model (HMM) [9] to model the actions.

As far as our survey and knowledge go, our proposal is the first framework for a complete human action discovery, learning and recognition that takes into consideration all aspects and parameters required to implement human action recognition in regular human living environment. Our work is also the first unsupervised approach in human action recognition using depth sensor data.

Taking advantage of the depth sensor, our framework uses simple learning algorithms. We note that the use of the depth sensor offers another advantage that privacy intrusion is partially reduced if only depth sensor data is being recorded and used.

3. Proposed Unsupervised Human Action Discovery, Learning and Recognition Framework

A complete human action recognition framework should comprise of at least three stages as shown in Fig. 1: discovery, learning and recognition. In addition, an observation module is required to segment and collect action instances from the sensor.

3.1 Overview of The Framework Fig. 2 shows our proposed framework to autonomously discover, learn and recognize human actions. The intelligent system continuously observes the person of interest and collect samples of the actions performed throughout a day or a specified period of time. The system performs the process of discovery and learning the models during idle time when the person is not available for observation, for example when the person is not at home or is sleeping.

The framework starts with capturing the sensor data. Predened set of features are extracted from the data and the continuous data is sampled or segmented into action instances. Each new action instance is checked against existing action models. Unrecognized action instances are collected into ac-

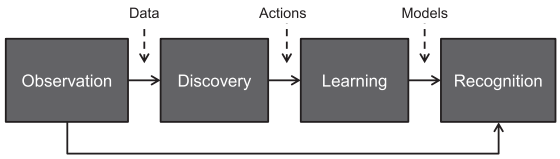

Fig. 1. Stages in Human Action Recognition.

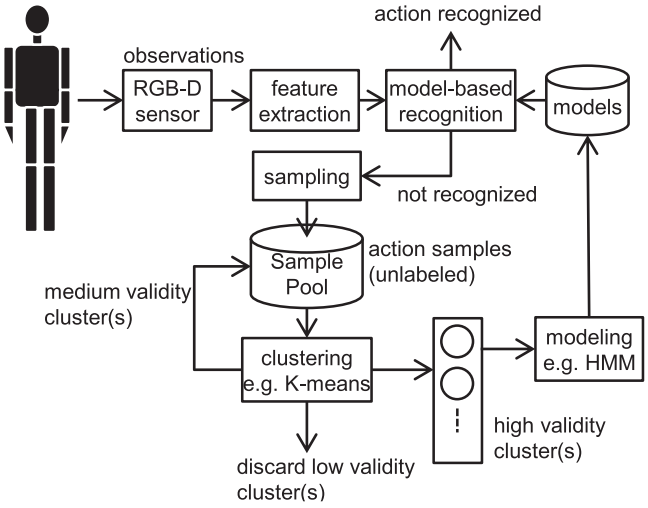

Fig. 2. Proposed Unsupervised Human Action Discover, Learn and Recognition Framework.

tion sample pool for the discovery of new actions from the pool. If there is no existing action model, as at the initial use of the framework, all action instances will be collected. This process of sample collection, discovery and learning is repeated continuously throughout the time that the intelligent system is with the person of interest. The learned models of actions are used to recognize new observations in real time.

We will elaborate the details of each step in the following subsections.

3.2 Feature Extraction and Sampling The features of the input data to the framework are 3-D coordinates of local vectors representing the human range of movement (ROM) [11]. The features are transformed and scaled to ensure view-invariant and scale-invariant.

Studies in kinematics of human motion [10] have identified that human motions are constrained within a range of movement (ROM): exion, extension, lateral exion, rotation of spinal column (the body movements); exion, extension, abduction, adduction of shoulder joint (the arm movements); exion, extension of elbow joint (the forearm movements); exion, extension of knee joint (the leg movements); exion, extension, adduction of hip joints (the thigh movements). We use this domain information to define the features for representation of human actions to enable discovery and recognition of all possible actions. In contrary, if the features were selected from a given dataset, the features would be fit to the current dataset and would not allow discovery of unseen actions. For example, if the dataset has 10 actions involving only hands. The features selected based on this dataset, for example by correlation based feature selection, will not allow discovery of new actions that will involve leg movements.

There are three typical ways to sample human actions or activity observations: manually identify start and end of every action, automatic detection of start and end of every action and, sliding window. Many research works analyze an entire video clip from standard dataset, where the start and end of every action or activity have been manually defined. Manual process is not only laborious but also not feasible in the regular human living environment. Automatic detection is usually achieved by detecting significant changes in the movement. However in everyday life, people spend significant amount of time in stationary state such as sitting and standing, which in our context are considered actions. Further, many actions are not carried out abruptly and such actions will not easily trigger automatic detection. Automatic detection remains a challenging task.

For the use of HAR in our homes, it is useful to have continuous observation of our activities at home. The sliding window is a suitable choice in our framework. Using fixed width keeps the operation simple. Schindler and Gool [12] have shown that 5 to 7 frames are sufficient to recognize basic actions. Sung *et al.* [13] have used fixed window of three seconds on the same dataset that we are using in this work. The choice of the window width can vary depending on the types of actions. Unlike sport activities, actions performed in daily activities at home are expected to be slow. For example, the normal stride frequency of leg movements during human locomotion is between 0.83-1.95 Hz [14]. In this work, we have used a window width of two seconds to represent an action instance. Based on the above mentioned stride frequency, we have made the assumption that around two cycles of human limb movement can be captured in two seconds.

```
inc_discovery( X, MinPt ):
    n = size of X;
    k = sqrt( n );
    Set the value of MinPt;
    While k ≥ 2 do:
        Cluster X into k clusters;
        Evaluate the homogeneity of each cluster that
        has at MinPt members;
        Collect the most homogeneous cluster, C*;
        Remove the members of C* from X, i.e.
        X = X − C*;
        Compute new k = sqrt( n* ) where n* is the size
        of trimmed X;
```

Fig. 3. The algorithm to incrementally discover activities with *MinPt* parameter [4].

3.3 Action Discovery: Clustering The key to action discovery is the ability to distinguish the various potential new actions within the unknown observations. Action discovery makes use of an incremental approach of clustering [4] as shown in Fig. 3. Conceptually, the algorithm is executed at regular intervals of time after a set of sufficient observations has been added into a sample pool; say, at the end of each day.

The algorithm does not have information with regards to the potential number of actions in the sample pool. The algorithm does not attempt to discover all actions within the sample pool at once. It makes its best effort to discover the most likely action in an incremental manner. By most likely action, the algorithm finds the most compact cluster using a mean variance $\bar{\sigma}^2$ measure given in Eq. 1. Low value of mean variance indicates compactness of the cluster.

$$mean\ variance\ \overline{\sigma}^2\left(C_j\right) = mean\left(var\left(C_j\right)\right) \cdots\cdots\cdots (1)$$

$$var\left(C_j\right) = \frac{1}{n_j - 1}\sum_{x_i \in C_j}\left(x_i^{(j)} - \overline{x}^{(j)}\right)^{\wedge 2} \cdots\cdots\cdots\cdots (2)$$

$$centroid\ \overline{x}^{(j)} = mean\left(x_i \in C_j\right) \cdots\cdots\cdots\cdots\cdots (3)$$

where $x_i^{(j)} = [x_{i1} \cdots x_{ip}]$ is a data point in Cluster C_j with dimension p, n_j is the number of points in Cluster C_j, $\bar{x}^{(j)}$ (dimension p) is the mean of all points in Cluster C_j, $\wedge 2$ is element-wise square.

3.4 Action Learning: Probability-Based Model

When a new action is discovered, i.e., a cluster is found, members of the cluster become examples to learn a template or model for that action. Fig. 4 summarizes the process of the learning phase. At the beginning of the learning phase, the cluster members are split into training and cross-validation sets. We use 8:2 training to cross-validation ratio. 80% of the members in the cluster are used to train the model, while 20% are used to perform holdout cross-validation of the model. Through exhaustive search of model parameters, the model with best cross-validation performance is chosen as the learned model.

In this work,we chose Hidden Markov Models (HMM) [6] to model the actions. An HMM is characterized by the fol-

lowing parameters:

$$\lambda = (A, B, \pi) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (4)$$

where

$A = \left\{a_{ij}\right\},\ 1 \le i, j \le N$ is the state transition matrix, which defines probability that the next state will be S_j given current state is S_i, where $S = \{S_i\},\ 1 \le i \le N$ is the set of states and N is the number of states.

$\pi = \{\pi_i\},\ 1 \le i \le N$ is the initial probabilities, where π_i is the probability of starting in State S_i.

$B = \left\{b_j(O)\right\},\ 1 \le j \le N$ is the emission probability matrix, where $b_j(O)$ describes the probability density of observing a continuous output O in state S_j.

Our framework is designed to discover, learn and recognize potentially undefined number of actions. It will significantly degrade the quality of the model if the continuous output is quantized into a set of finite discrete output symbols. The required number of the output symbols will be prohibitively large in order to cover the 3-D volume around a human body while cater for the required resolution to discriminate similar actions. For this reason, Mixture of Gaussians is used to represent the output observations. For M-component Mixture of Gaussians, the emission probability density function is given as

$$b_j(O) = \sum_{k=1}^{M} c_{jk} \mathcal{N}\left(x, \mu_{jk}, \Sigma_{jk}\right) \cdots\cdots\cdots\cdots\cdots\cdots (5)$$

where c_{jk} is the mixture weight, $\mathcal{N}$ is the Gaussian or normal density function, μ_{jk} and Σ_{jk} are the mean vector and covariance matrix associated with state S_j and mixture component k.

The parameters to be learned for the HMM model are a_{ij}, π_i, c_{jk}, μ_{jk} and Σ_{jk}. The parameters are estimated using the Baum-Welch algorithm [6]. The description of the update algorithm is beyond the scope of this paper, but good references for it can be found in Rabiner [6], Murphy [9] and Yang *et al.* [15]. To autonomously determine the values of the number of states N and the number of mixture of Gaussians M, the algorithm uses an exhaustive search within a range of values and select the model that has the best performance in cross-validation test. Exhaustive search is however computationally demanding. It is necessary to constrain the range of the search values. We will suggest a suitable range of values based on our experiment outcomes.

ALGORITHM: Learning of action model

INPUT: A cluster of observations $C = \{O_1, \ldots, O_n\}$

OUTPUT: HMM model of the action $\lambda = (A, \pi, c, \mu, \Sigma)$

ALGORITHM:

1: Set values of Q_{min}, Q_{max}, M_{min}, M_{max} and P_{train};
2: LL = -Inf;
3: Split members of C into training set C_{train} and cross-validation set C_{cv} with $|C_{train}| = P_{train}|C|$ and $|C_{cv}| = (1 - P_{train})|C|$;
4: For M_{min} to M_{max} %search number of mixtures
  4.1: For Q_{min} to Q_{max}; %search number of states
    4.1.1: Train HMM on C_{train} to give $\hat{\lambda} = (\hat{A}, \hat{\pi}, \hat{c}, \hat{\mu}, \hat{\Sigma})$;
    4.1.2: Test the HMM on C_{cv} and computer average log likelihood LL_{cv};
    4.1.3: If $LL_{cv} > LL$, $\lambda = \hat{\lambda}$;

Fig. 4. The steps involved in learning phase. Q_{min} and Q_{max} are the minimum and maximum value of the number of states respectively. M_{min} and M_{max} are the minimum and maximum value of the number of mixtures respectively. P_{train} is the proportion of data to be used as training set.

4. Data and Experiment

4.1 Data In this work, we have used the dataset "Cornell Activity Dataset CAD-60" [16] to demonstrate the effectiveness of our proposed framework. CAD-60 is the earliest published human activities dataset based on Kinect data. CAD-60 consists of twelve daily activities by four subjects (Person 1 to 4). Still (standing) and random activity samples by each subject are also included in the dataset. We use only the skeleton data in the dataset to obtain the required features as described in Section 3.

Including still, CAD-60 dataset has thirteen activities in total. Among them, four of the activities have less than ten examples per subject and are insufficient for the discovery phase. We have therefore used the remaining nine activities and the random activities to test our framework. These activities are atomic, i.e., actions in our context, and are listed in Table 1. The labels in Table 1 are used for the purpose of evaluation of the effectiveness of the framework. The labels are not provided to the framework during runtime.

4.2 Experiment We sampled 80 instances of each action for each subject using sliding window of two seconds. We sampled 160 instances from the random actions for each subject. The samples are split into learning and test set at 7:3 ratio. For each subject, we have a learning set of 56 instances of 9 actions in addition to 112 instances of random movements. The instances from different actions are mixed together and the framework does not know the labels. This gives a total of $56 \times 9 + 112 = 616$ instances in the learning set, which is equivalent to the sample pool in Fig. 2. Likewise, we have a total of $24 \times 9 + 48$ instances mixed together without label in the test set, which is equivalent to the unseen observations to be recognized.

The learning sets are used for discovery phase, which the outcome is a set of clusters to be fed into the model learning phase. Since the framework do not have labels of the observations, it will learn a model for each cluster irrespective of the purity of the cluster. The test sets are used to evaluate the recognition ability of the learned model.

In our experiment, we let the algorithms in the framework exhaust all samples in each dataset. The datasets have finite samples. When implemented in a real life situation, the algorithms in the framework do not have to discover all actions at once. In our experiment, we have used minimum points per

Table 1. List of actions

1.	A1	Brushing teeth
2.	A2	Cooking (chopping)
3.	A3	Cooking (stirring)
4.	A4	Relaxing on couch
5.	A5	Still (standing)
6.	A6	Talking on couch (sitting)
7.	A7	Talking on the phone
8.	A8	Working on computer
9.	A9	Writing on whiteboard
10.	RA	Random

Table 2. Precision and recall in evaluation of the framework. The results are given for data from each of the four subjects (Person 1 to 4) and average across the four subjects.

	Person 1		**Person 2**		**Person 3**		**Person 4**		**Average**	
	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
brushing teeth	100	58.3	100	62.5	100	83.3	100	29.2	100	58.3
cooking (chopping)	100	70.8	88.0	91.7	100	83.3	72.7	100	90.2	86.5
cooking (stirring)	96.0	100	100	79.2	100	95.8	100	62.5	99.0	84.4
relaxing on couch	92.0	95.8	100	100	85.2	95.8	100	95.8	94.3	96.9
still (standing)	96.0	100	100	95.8	100	100	80.0	100	94.0	99.0
talking on couch	77.4	100	100	83.3	100	100	73.3	91.7	87.7	93.8
talking on the phone	96.0	100	70.0	87.5	100	91.7	92.3	100	89.6	94.8
working on computer	100	87.5	100	100	100	91.7	69.0	83.3	92.2	90.6
writing on whiteboard	100	100	100	100	64.9	100	100	66.7	91.2	91.7
Average:	**95.3**	**90.3**	**95.3**	**88.9**	**94.5**	**93.5**	**87.5**	**81.0**	**93.1**	**88.4**

cluster $MinPt = 25$, minimum number of mixtures $M_{min} = 1$, maximum number of mixtures $M_{max} = 7$, minimum number of states $Q_{min} = 2$ and maximum number of states $Q_{max} = 15$. Given that we have 15 frames per feature in one sample, we set $Q_{max} = 15$. We would like to evaluate the learning algorithm from lowest possible Q value.

5. Results and Discussions

To evaluate the performance of the framework, we test the recognition ability of the learned models on unseen observations, i.e., the test sets as described in Section 4. The results of the evaluation is summarized in Fig. 5. Table 2 gives the detail result. The framework achieved a recognition rate of an average precision of 93.1% and recall of 88.4%. The low recall rate is due to the poor performance in brushing teeth action (A1).

From Table 2, we can see that the low recall rate of the brushing teeth action is largely due to the poor performance in the data of Person 1 and Person 4. Apart from these two cases, the performance to recognize the other actions is good.

Further to the evaluation of recognition rate, we have observed that suitable values of M and Q are in the range of $1 \leq M \leq 4$ and $2 \leq Q \leq 6$. This knowledge will allow us to reduce computation time of the learning phase.

As we have not come across unsupervised learning using the skeleton data from RGB-D sensor, we compare our results with the work of the authors of CAD-60 [13], and a latest result from Wang *et al.* [17], that were both based on supervised learning using the same dataset. Their experiments involved detecting activities based on locations such as kitchen and living room. By zoning the activities, their algorithms were required to discriminate at most four activities and the random activities at any one time. In contrast, our results were obtained through an unsupervised approach from discovery to recognition. Our algorithms were required to discriminate 9 actions and the random action at any one time in both discovery and recognition phases.

We compute the average precision and recall of the results of Sung *et al.* [13] for the 9 activities we have used in our experiment, and compare with our results in Table 3. Since "talking on phone" was evaluated in three locations (bedroom, living room and office) in their results, we have taken the average value for the precision and recall. The numbers for their results are summarized from their Full Model "Have Seen" results. For the results from Wang *et al.* [17], we computed the precision and recall from their confusion matrix on "same-person" setting.

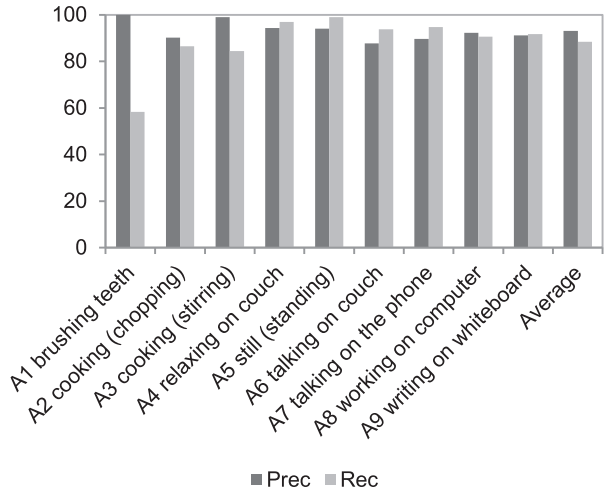

Fig. 5. Average precision and recall across all four subjects (Person 1 to 4) for each activity in evaluation of the framework.

Table 3. A comparison of our results and the results of Sung *et al.* [13].

	This paper		**Sung *et al.*** [13]		**Wang *et al.*** [17]	
	Prec	**Rec**	**Prec**	**Rec**	**Prec**	**Rec**
brushing teeth	100	58.3	96.7	77.1	71.4	83.3
cooking chopping	90.2	86.5	70.3	85.7	100	100
cooking stirring	99.0	84.4	74.3	47.3	100	83.3
relaxing on couch	94.3	96.9	86.8	82.7	100	100
talking on couch	87.7	93.8	98.8	94.7	100	100
Still	94.0	99.0	NA	NA	100	100
talking on phone	89.6	94.8	88.4	91.1	100	100
working on computer	92.2	90.6	89.5	93.8	100	100
writing on whiteboard	91.2	91.7	85.5	91.9	100	100
Average	**93.1**	**87.3**	**86.3**	**83.0**	**96.8**	**96.3**

Our results are on average superior to that of Sung *et al.* [13], however the results of Wang *et al.* [17] is almost perfect for seen person. We restate that both of them have used supervised approach, while our results were obtained without giving the labels to our framework.

6. Conclusions

We have proposed an autonomous human action discovery, learning and recognition framework that takes unlabeled skeleton data from an inexpensive depth sensor. We have

shown the effectiveness of the framework on a third party dataset of human daily activities. Our results show the performance of our unsupervised approach is in-par with another work using supervised approach on the same dataset. The framework achieved an average precision of 93.1% and recall of 88.4% in recognition of unseen observations.

Based on our empirical result, it is sufficient to constrain the parameters of the HMM models, M and Q, in the range of $1 \leq M \leq 4$ and $2 \leq Q \leq 6$. The fact that the framework uses unlabeled data from an inexpensive depth sensor will allow it to be used in regular human living environment without requiring expensive modification to the environment, wearing of sensors and human labeling of the data. The use of depth data also helps moderate the concern of privacy in vision-based monitoring.

References

(1) P. Turaga, R. Chellappa, V.S. Subrahmanian, and O. Udrea: "Machine recognition of human activities A survey", IEEE Trans. on Circuits and Systems for Video Technology (2008)

(2) R. Poppe: "A survey on vision-based human action recognition", Image and vision computing (2010)

(3) J.K. Aggarwal and M.S. Ryoo: "Human activity analysis A review", ACM Computing Surveys (2011)

(4) W.H. Ong, T. Koseki, and L. Palafox: "An Incremental Approach of Clustering for Human Activity Discovery", IEEJ Trans. EIS (2014)

(5) E. Kim, S. Helal, and D. Cook, Diane: "Human activity recognition and pattern discovery", IEEE Pervasive Computing (2010)

(6) L.R. Rabiner: "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE (1989)

(7) J. Yamato, J. Ohya, and K. Ishii: "Recognizing human action in time-sequential images using hidden Markov model", IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) (1992)

(8) L. Xia, C.C. Chen, and J.K. Aggarwal: "View invariant human action recognition using histograms of 3D joints", IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2012)

(9) K.P. Murphy: Markov and hidden Markov models, Machine learning: a probabilistic perspective (2012)

(10) V.M. Zatsiorsky: Kinematics of human motion, Human Kinetics (1998)

(11) W.H. Ong, T. Koseki, and L. Palafox: "Unsupervised Human Activity Detection with Skeleton Data From RGB-D Sensor", 2013 Fifth International Conference on Computational Intelligence, Communication Systems and Networks (2013)

(12) K. Schindler and L. Van Gool: "Action snippets: How many frames does human action recognition require?", IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2008)

(13) J. Sung, C. Ponce, B. Selman, and A. Saxena: "Unstructured human activity detection from rgbd images", IEEE International Conference on Robotics and Automation (ICRA) (2012)

(14) J. Nilsson and A. Thorstensson: Adaptability in frequency and amplitude of leg movements during human locomotion at different speeds (1987)

(15) J. Yang, Y. Xu, and C.S. Chen: "Human action learning via hidden Markov model", IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans (1997)

(16) J. Sung, C. Ponce, B. Selman, and A. Saxena: "Human Activity Detection from RGBD Images", AAAI workshop on Plan Activity and Intent Recognition (2011)

(17) J. Wang, Z. Liu, and Y. Wu: Human Action Recognition with Depth Cameras (2014)

Wee-Hong Ong (Non-member) received the B.Eng. in Communication and Control Engineering from the University of Manchester, Institute of Science and Technology in 1997. He received the M.Sc. in Computing Science from the Imperial College London in 2004. He received the Doctor of Engineering in Electrical and Information Systems from the University of Tokyo, Japan. He is currently a lecturer in the Universiti Brunei Darussalam. His research interests are in personal robotic and ambient intelligence.

Leon Palafox (Non-member) received the B.Sc. degree in Electronic Engineering from the National Autonomous University of Mexico (UNAM). He received the M.Sc. degree and PhD degree in Electronic Engineering from the Graduate School of Engineering, The University of Tokyo, Tokyo, Japan. He was a Postdoctoral Fellow at UCLA, School of Medicine. He is currently a Postdoctoral Fellow in the University of Arizona. His research interests include neuroscience, machine learning, bioinformatics and planetary sciences.

Takafumi Koseki (Member) received the Ph.D. degree in electrical engineering from the University of Tokyo, Tokyo, Japan in 1992. He is currently a Professor in the Department of Electrical Engineering and Information Systems, School of Engineering, The University of Tokyo. His current research interests include application s of electrical engineering to public transport systems, especially to linear drives, and analysis and control of traction systems. Dr. Koseki is a member of the Institute of Electrical Engineers of Japan, the Institute of Electric and Electronic Engineers, the Japan Society of Mechanical Engineering, the Japan Society of Applied Electromagnetics and Mechanics, the Japan Society of Precision Engineering, and Japan Railway Electrical Engineering Association.